Ethan Tom, Tulasa Chitrakar, Jeffrey Lai

CIS 3920 ETRA[38217]

Term Project 22'

Professor Mohammad Chowdhury

Early Diabetes Indication

Introduction

Diabetes is among the most common chronic diseases in the United States, increasing yearly throughout the world and costing the world economy hundreds of billions. For those that are unfamiliar, diabetes is a chronic health condition in which people lose their ability to maintain blood glucose levels, resulting in a decreased quality of life and life expectancy. There is Type 1 diabetes, the body does not create insulin and Type 2, where the body makes insufficient insulin or does not utilize the insulin effectively. Glucose is released from the food we ingest and the pancreas produces insulin. The insulin aids the glucose to cells to provide energy. While there is no cure for diabetes, people can regulate their glucose by eating healthy, staying active, and receiving medical treatment. Predictive models for diabetes risk are crucial tools for the public and public health officials since early diagnosis can lead to lifestyle changes and more effective treatment.

Project Description

This project aims to detect diabetes through its key indicators and features. Having high blood glucose levels over time may severely cause health problems including heart disease, nerve disease, eye disease, blindness, kidney failure, foot sores and amputations. Being able to predict or early detect a serious medical condition can vitally alter one's life for the better. It determines a problem and allows the user to determine the best course of action.

Dataset Description

The dataset used is diabetes _ 012 _ health _ indicators _ BRFSS2015.csv, which can be found in https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. This dataset is composed of 22 features and 253680 records.The attribute which we are going to analyze and predict is Diabetes_012 that has 3 different classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. What makes this dataset easy to digest is the simple binary classification applied to the dataset. It is surveyed to provide boolean answers represented as a double data type. Early detection of diabetes includes such features but are not limited to:

- High blood pressure
- High cholesterol
- Smoking
- Age
- Sex
- Race
- Dieting
- Exercise
- Alcohol consumption
- (BMI) Body Mass Index

Literature Survey

There have been many studies done to see how people get diabetes. Again, there are two types of diabetes, type 1 and type 2, one of which is linked to genetics while the other is a disease that can be obtained more generally. Those cases also focus in particular on factors that cause diabetes in people. Most of the studies also focus on numbers such as age and more so lifestyle factors such as how sedentary a person is. There is also a focus on the biology of the person particular to their glucose tolerance and their insulin resistance. Most studies find that women that are age 25 or older are at higher risk of diabetes. You are also at higher risk depending on your physical activity. Most of the data is based on worldwide data opposed to location based, however there is data that shows which country is the most diabetic country also which country has the highest percentage of diabetes.

Project Methodology

1. Looked through the data to see if there were any null values within the data.
2. Explored the data to find it had 20 different criteria for characterizing if someone has diabetes.
3. Given the data's large range of characteristics, we narrowed down on some features to focus our project on: age, BMI, mental health, and physical health.
4. Identified and plotted the relationship of these traits against diabetes status using a box and whisker plot.
5. Split data into training and testing sets using a 70/30 split.
6. Trained a KNN classifier and performed the prediction and viewed its results. Used a confusion matrix to check the accuracy.
7. Created a naive bayes classifier for the data. Check it using a confusion matrix to check for accuracy.
8. Train the data again and create a decision tree, checking for diabetes. Use a confusion matrix after to check accuracy.

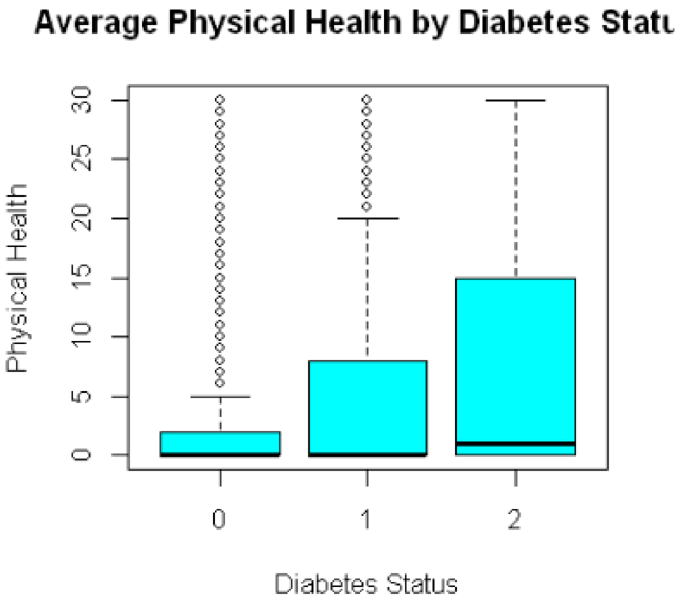
Analysis / Results

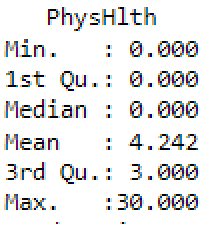
The series of box plots shows the relationship between:

- Average Age vs Diabetes
- Average BMI vs Diabetes
- Average Mental Health vs Diabetes

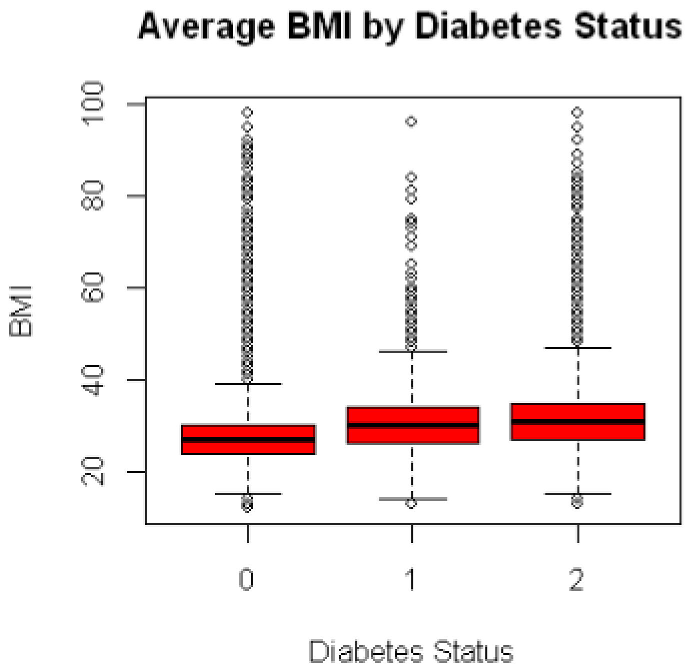
- Average Physical Health vs Diabetes

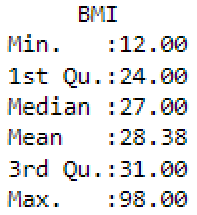
Physical Health vs Diabetes: has a lot of outliers with the max value of 30.00 and the mean of 4.242 as shown in the boxplot. Most of the outliers fall into class 0, while there are much fewer outliers in class 2. Similarly, Mental Health vs Diabetes; All the classes have just about an equal amount of outliers. While the median = 0.00 and the mean = 3.185, the max value for mental health = 30.00. Average BMI: mean = 28.38 and 3rd quartile = 31.00, while max = 98.00. The box plot displays how far off the data points for average bmi and diabetes are. While outliers can include errors, they can sometimes contain useful information about any underlying system. The Average Age: mean = 8.032, median = 8.00, and the max = 13.00. There are fewer outliers in class 2, but overall this has less error and could give more accurate information.



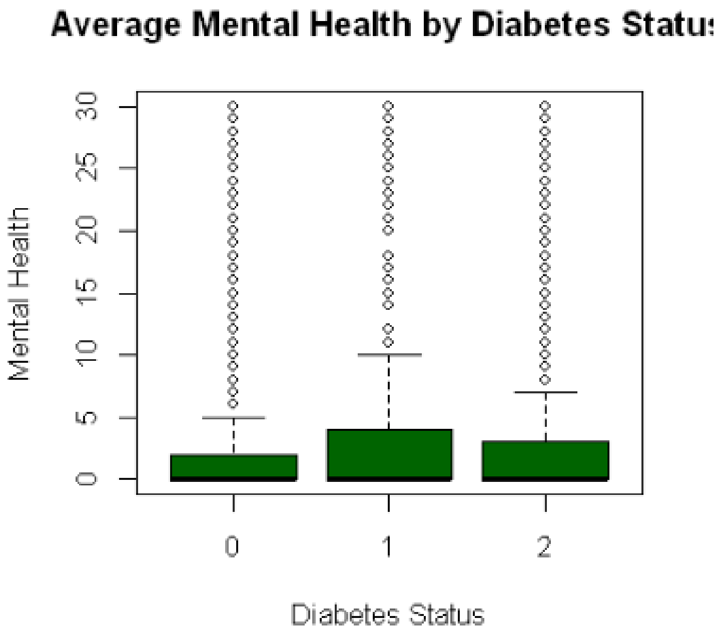


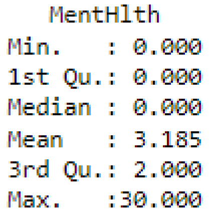
```
   PhysHlth
Min.   : 0.000
1st Qu.: 0.000
Median : 0.000
Mean   : 4.242
3rd Qu.: 3.000
Max.   :30.000
```



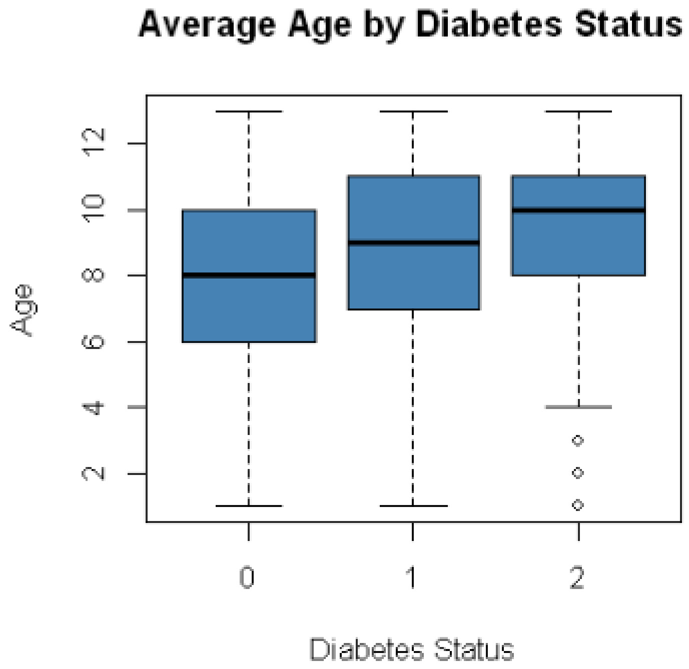


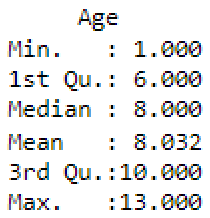
```
       BMI
Min.   :12.00
1st Qu.:24.00
Median :27.00
Mean   :28.38
3rd Qu.:31.00
Max.   :98.00
```





```
    MentHlth
Min.   : 0.000
1st Qu.: 0.000
Median : 0.000
Mean   : 3.185
3rd Qu.: 2.000
Max.   :30.000
```





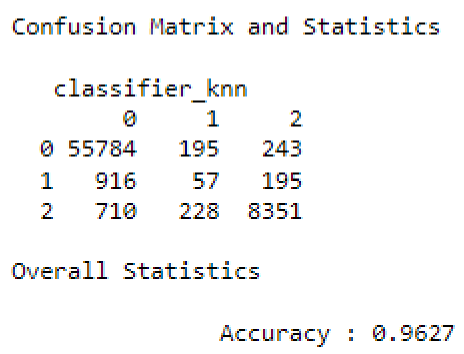
```
      Age
Min.   : 1.000
1st Qu.: 6.000
Median : 8.000
Mean   : 8.032
3rd Qu.:10.000
Max.   :13.000
```

The KNN method is a non-parametric supervised learning approach, which means it makes no assumptions about the mapping function's structure. The non-parametric function aids in the ability to generalize to previously unseen data. Because the KNN method is based on the distance between data points, it produces predictions for new data based on the k most comparable training patterns.

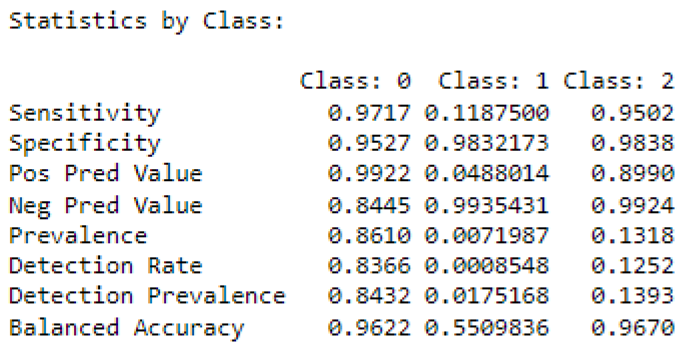
For training and testing the data, we used a 70/30 split on the dataset. 70% of the data would be used to train our model while the remaining 30% would be used to test our model. Before splitting, we removed any empty data from our dataset. After splitting the data by rows, we used the KNN classification for prediction where the k (number of nearest neighbors) is equal to 2.

Confusion matrix is a table that is used to define the performance of a classification algorithm. It gives us the actual and predicted values. After we did a KNN algorithm testing on the diabetes column, we can see the actual values as well as the predicted values, for all the classes in that column. This algorithm provides 96.27%

accuracy. The confusion matrix also shows the sensitivity rate for all the classes as well as the accuracy on the actual and predicted values among those classes.

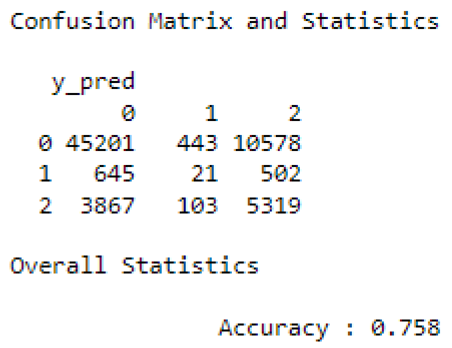


```
Confusion Matrix and Statistics

   classifier_knn
        0     1     2
 0 55784   195   243
 1   916    57   195
 2   710   228  8351

Overall Statistics

                Accuracy : 0.9627
```

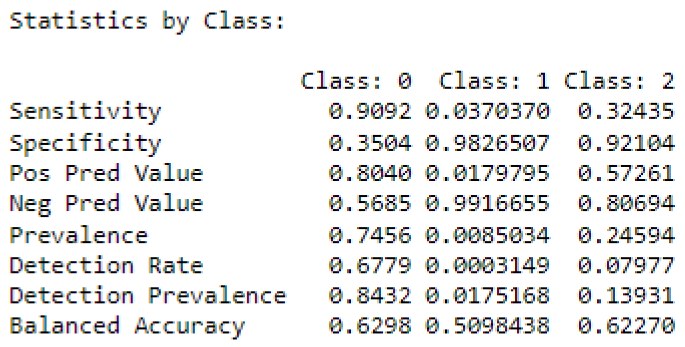


```
Statistics by Class:

                     Class: 0  Class: 1 Class: 2
Sensitivity            0.9717 0.1187500   0.9502
Specificity            0.9527 0.9832173   0.9838
Pos Pred Value         0.9922 0.0488014   0.8990
Neg Pred Value         0.8445 0.9935431   0.9924
Prevalence             0.8610 0.0071987   0.1318
Detection Rate         0.8366 0.0008548   0.1252
Detection Prevalence   0.8432 0.0175168   0.1393
Balanced Accuracy      0.9622 0.5509836   0.9670
```

In the Naïve Bayes classification model, we assume that the attributes are independent of one another. We use the same training and testing model as the one we used for KNN. Bayes classification is based on Bayes' theorem that finds the probability of an event occurring given the probability of another event that has already occurred. The assumptions made in this model are generally not correct, which is why the accuracy for this model is 75.8%, which is 20.47% less accurate than the KNN model. Also, the accuracy among the class is relatively lower. While Bayes classification model is less accurate, it is a fast and easy method to predict a test data set and works well in multi-class predictions. Also, because our attributes are categorical, this model performs better than other models.

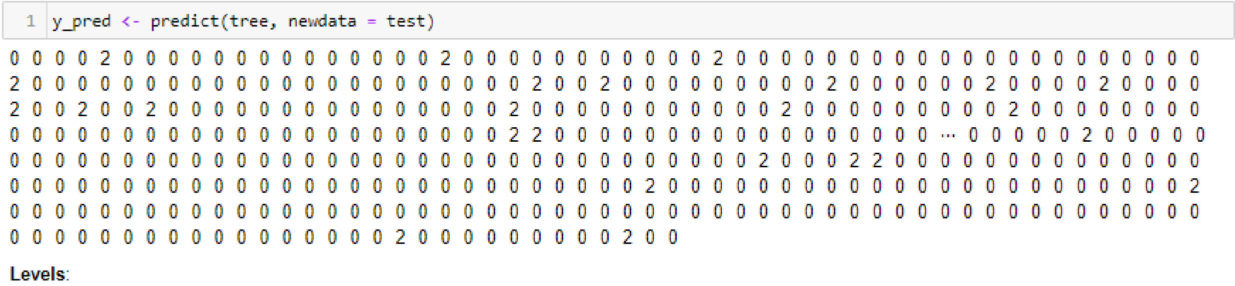


```
Confusion Matrix and Statistics

   y_pred
        0     1     2
 0 45201   443 10578
 1   645    21   502
 2  3867   103  5319

Overall Statistics

                Accuracy : 0.758
```



```
Statistics by Class:

                     Class: 0  Class: 1 Class: 2
Sensitivity            0.9092 0.0370370  0.32435
Specificity            0.3504 0.9826507  0.92104
Pos Pred Value         0.8040 0.0179795  0.57261
Neg Pred Value         0.5685 0.9916655  0.80694
Prevalence             0.7456 0.0085034  0.24594
Detection Rate         0.6779 0.0003149  0.07977
Detection Prevalence   0.8432 0.0175168  0.13931
Balanced Accuracy      0.6298 0.5098438  0.62270
```

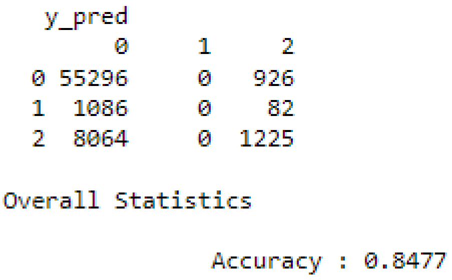
Using the same training and testing dataset, we used the decision tree model to train our model and get a prediction for the test datasets. We can see that the predicted results fall mostly on class 0 and class 2.



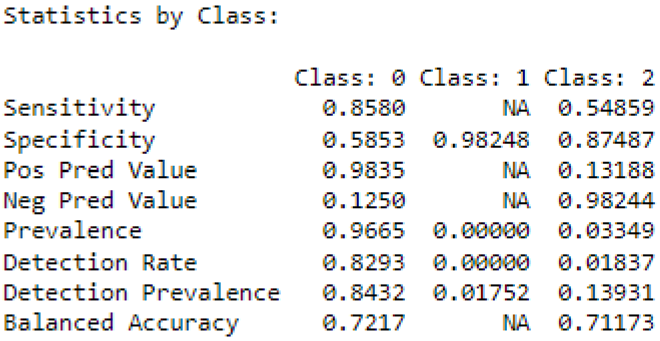
```
y_pred <- predict(tree, newdata = test)
```

```
0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 2 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 2 0 0 0 0 2 0 0 0 0
2 0 0 2 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ⋯ 0 0 0 0 0 2 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 2 0 0
Levels:
```

Confusion matrix also summarizes that for us. We can see that the predicted values either fall on class 0 or class 2. None of the predicted values belong to class 1. It seems a bit unusual, which is why the accuracy of this model is only 84.77%. Also, the sensitivity for class 1 is NA, because none of the predicted values fall on that class. Although the overall accuracy of this model is 84.77%, the accuracy by class is moderately low.



```
   y_pred
        0     1     2
  0 55296     0   926
  1  1086     0    82
  2  8064     0  1225

Overall Statistics

               Accuracy : 0.8477
```



```
Statistics by Class:

                     Class: 0 Class: 1 Class: 2
Sensitivity            0.8580       NA  0.54859
Specificity            0.5853  0.98248  0.87487
Pos Pred Value         0.9835       NA  0.13188
Neg Pred Value         0.1250       NA  0.98244
Prevalence             0.9665  0.00000  0.03349
Detection Rate         0.8293  0.00000  0.01837
Detection Prevalence   0.8432  0.01752  0.13931
Balanced Accuracy      0.7217       NA  0.71173
```

Summary / Conclusion

Diabetes can cause damage to your nerves that control your heart and blood vessels. This damage can lead to heart disease over time. Utilizing the box and whisker plot helps viewers visualize the data for simplified consumption of the data. We had decided to isolate just a few traits for simplicity sake to visualize part of the data in regards to some more general sets. The KNN model was a great algorithm to classify the data. Using k = 2 labeled the data into its nearest corresponding classes, predicting with an accuracy of 96.27% according to the confusion matrix we used to validate the model's accuracy. The decision tree unfortunately did not fare so well. The decision tree was not an accurate model considering the large number of features considered to determine a person's diabetes status, ringing in at an accuracy of 84.77%. Thinking logically about the efficacy of a decision tree in context to something like diabetes status, it makes sense as there is not a set format to which we can say that a given set of traits will result in someone having diabetes. At an even lower clip of accuracy, the naive bayes model came in at 75.8% accuracy. Something that might contribute to the inaccuracy is the assumption of independence between the features that is used to create the model. Though there is shared categorical data and such, there are features that are related in one way or another in our data set, such as cholesterol levels and blood pressure.

Sources

- Machine Learning Fundamentals: The Confusion Matrix
- Diabetes Health Indicators Dataset | Kaggle.
- Confusion Matrix - an overview | ScienceDirect Topics.